一种星系形态分类的新方法-GMC *

王林倩[1]，邱波[1+]，罗阿理[2]，孔啸[2]，逯亚坤[1]，郭小雨[1]

(1.河北工业大学，天津 300401；2. 中国科学院国家天文台，北京 100012)

摘 要：在天文学研究领域，星系的分类一直是一个热点和难点问题。近年来有学者将机器学习应用到星系形态的简单分类任务上，但在分类过程中出现特征选择困难、特征遗漏、分类器选择困难等一系列难题。星系在视觉形态上可以分为椭圆星系、旋涡星系、透镜星系以及不规则星系，本文针对 SDSS DR16、Galaxy Zoo2、EFIGI 目录中星系的测光图像，提出了一种分类精度更高的星系形态分类的方法 GMC(Galaxy Morphological Classification)。我们首先对图像进行了剪裁、去噪处理，然后采用旋转、平移、缩放等方法进行数据增强，最后搭建了星系形态分类网络 GMC-net 对图像进行分类。从实验分类结果来看，旋涡星系、椭圆星系、透镜星系以及不规则星系分类精确率分别为 98.29%、98.49%、99.18%、99.91%，召回率分别为 98.44%、99.03%、98.89%、99.34%；对单独来自 EFIGI 目录中四种形态星系的分类准确率也达到了 99.34%。实验结果表明 GMC 相较于其他分类方法表现更好，可以更有效地用于星系的形态分类。

关键词：星系形态分类；数据增强； 卷积神经网络；

1 引 言

随着观测技术的进步、天文观测仪器的发展，大型数字巡天计划如斯隆数字巡天(Sloan Digital Sky Survey，SDSS[1]，COSMOS巡天(Cosmic Evolution Survey，COSMOS)[2]，大口径全天巡视望远镜LSST(Large Synoptic Survey Telescope，LSST)[3]等逐步实施，星系观测数据呈现出爆炸式增长的趋势。

星系是众多天体中一类，主要由恒星、恒星遗骸、星际气体、尘埃和暗物质等组成，并受引力绑定的一个系统。星系的形态与星系的形成、演化有着密切的联系，是探究星系物理的重要参数。随着机器学习和深度学习在各个领域大放光彩，星系形态的自动分类方法也迅速发展。Freed M[4]用多个支持向量机(SVM)对星系形态进行螺旋星系、椭圆星系和不规则星系的三分类，其最高分类准确率为96.8%。Dieleman [5]等以 5 万多张星系图片为训练集，经过100多次的尝试，首次提出用卷积神经网络进行模型训练，最终以0.07492的RMS值获得了“银河动物园挑战赛”比赛的冠军。Kim et al[6]利用SDSS DR12中17344张恒星和47656张星系图像，提出一个类似VGG的11层深度卷积神经网络实现了对恒星、星系进行分类，测试集上的准确率值分别可以达到99.52%和99.48%。I. M. Selim[7]等对来自于EFIGI目录的旋涡星系、椭圆星系、 透镜星系和不规则星系进行了四分类，提取了星系图像的颜色特征、纹理特征以及其形状特征三种特征，并用二进制正弦余弦算法选择最相关的特征，最后用KNN对四类星系分类的准确率分别为97.43%、100%、79.48%、100%，平均分类准确率为94.2%。Ansh Mittal[8]等提出了一种星系形态的分类网络daMCOGCNN，该方法对不规则星系进行了数据增强、使用不同的激活函数构建了卷积神经网络，使椭圆星系、旋涡星系和不规则星系分类准确率达到了97%。Mittal[9]等结合数据增强技术和深度学习的方法实现了对透镜星系、

* 基金项目：国家自然科学基金委员会-中国科学院天文联合基金（U1931134）， 河北省自然科学基金（A2020202001），中国科学院天文大科学研究中心 LAMOST 重大成果培育项目

作者简介：王林倩，女，硕士.研究方向：机器学习，图像处理. E-mail：1989147094@qq.com

+通讯作者：邱波，男，博士生导师. 研究方向：机器学习，模式识别. E-mail：qiubo@hebut.edu.cn

椭圆星系和旋涡星系的分类，此模型的分类准确率达到了90.2%，验证准确率达到了88.3%。Hosny[10]等提取星系图像的非冗余色彩特征，并提出了一种寻找最优的特征子集方法，最后利用极端机器学习(EML)对椭圆星系、旋涡星系、透镜星系和不规则星系进行分类，分类效果达到98%。

然而，目前对于星系形态分类研究领域还存在分类类别少、分类样本类间比例失衡等问题，此前研究多是对椭圆星系、涡旋星系、透镜星系进行二分类或三分类。面对更多类型星系形态的数据，当前的分类方法所得的准确率比较低，因此迫切需要一种能准确区分更多星系形态的方法。我们的目标是能够找到一种方法能够实现旋涡星系、椭圆星系、透镜星系以及不规则星系自动分类，甚至可以实现不同数据库中四类不同形态星系的自动分类。如图1所示，本文对来自不同数据库中的星系图像进行了裁剪和下采样从而筛除质量差的数据，同时对数据进行去噪处理和数据增强来减小图像噪声和样本类间比例失衡对分类模型的影响。之后我们提出了一种更高效的星系形态自动化分类网络GMC-net，回避了图像特征提取、选择、分类器选择这些难题，从而高效的实现了四类不同形态星系的分类。

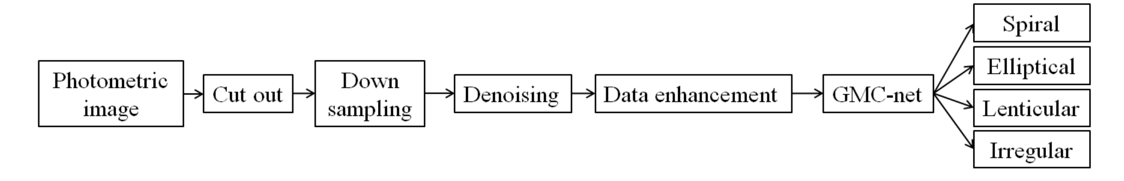


图 1　GMC整体流程图

Fig.1 The overall flow chart of GMC

2 数 据

本次研究主要使用的是SDSS DR16、Galaxy Zoo2和EFIGI目录的数据，从本质来说三个数据库的测光数据来源都是SDSS数字巡天。SDSS所得到的原始数据为u、g、r、i、z五个波段数据，但u、z波段多是近紫外和近红外，且包含的有用信息非常少。g，r，i三波段数据已经完全足够还原比较真实星系图像，所以当前相关研究一般采用 g，r，i三波段数据合成的图像[4]~[8]，[25][26]。

2.1 数据获取

EFIGI目录[13]中的测光和光谱数据是从SDSS DR5目录获得的，目录中星系按形态主要分为椭圆形、透镜状、旋涡形、不规则形、矮形（Dwarf），此五类又分为不同的子类。利用星系形态参数T(T ∈[-6，11]，T为整数，分别代表不同形态星系类型)可以筛选不同形态的星系，表1中展示了各类星系的选择标准，最终在EFIGI获得的星系为920张旋涡星系、289张椭圆星系、531张透镜星系以及248张不规则星系。

星系动物园(Galaxy Zoo2)[14]包括11个任务和37个响应，同一个样本超过二十人对其分类才会统计，[14]给出每个分类任务干净样本阈值范围以及11个具体分类任务，为保证所选样本更准确，此次设置的阈值均大于所建议阈值，表1注释部分对各个参数阈值设置进行了详细解释所示，最终在Galaxy Zoo2获得的星系为3095张旋涡星系、4208张椭圆星系、1805张透镜星系以及235张不规则星系。

本次研究采用了最新发布SDSS DR16[16]测光数据，该数据星表可在CasJobs[15]中通过星系specObjID与Galaxy星表交叉得到相应星系赤经赤纬。除了表1所叙述的主要查询标准限制，还有如下设置：所有图像都设置红移下限为0.001、红移上限为0.025、通量下限为50、通量上限为500及0.01的图像缩放因子，设置提取top2000个数据。不规则星系物理条件的限制目前还未知，在此未得到不规则星系。DR16中各类星系数量分布也是不均的，在此人工筛选去除了双重的、合并的以及包含许多未知对象的图像最终得到913张旋涡星系、1956张椭圆星系、805张透镜星系。

表 1 星系数据选择标准

Tab.1 Galaxy data selection criteria

	EFIGI	Galaxy Zoo2		SDSS DR16	
Class	Sample selection	Tasks	Threshold setting	Main query criteria	N_{sample}
Spiral	Sb(T=3) Scd(T=6)	T01 T02 T04	$f_{features/disk}$ >0.430 $f_{edge\text{-}on,no}$ >0.750 $f_{spiral,yes}$ > 0.719	*g.lnLDeV_g* < -2000.0 *g.lnLDeV_g* + 0.1 < *g.lnLExp_g*	4928
Elliptical	cE(T=-6) E(T=-5) cD(T=-4)	T01 T07	f_{smooth} >0.469 $f_{in_between}$ >0.70	*g.lnLDeV_r* >*g.lnLExp_r* + 0.1 *g.lnLExp_r* > -999.0 *g.lnLDeV_g* > -999.0	6453
Lenticular	$S0^-$(T=-3) $S0^0$(T=-2) $S0^+$(T=-1)	T01 T02	$f_{features/disk}$ >0.630 $f_{edge\text{-}on,yes}$>0.785	*g.lnLDeV_r* < *g.lnLExp_r* + 0.1 *g.lnLDeV_g* + 0.1 >*g.lnLExp_g* -1200.0<*g.lnLDeV_g* < -1500.0	3141
Irregular	Im(T=10)	T01 T02 T03 T04 T05 T06 T08	$f_{features/disk}$ >0.430 $f_{edge\text{-}on,no}$ >0.715 f_{no_bar} >0.715 $f_{spiral,no}$ >0.715 f_{No_bulge} >0.750 $f_{odd,yes}$ >0.650 $f_{irregular}$ > 0.715	-	483

注：EFIGI中样本选择中前面字母(例$S0^0$)为所代表的星系形态类型，括号中T为形态参数；Galaxy Zoo2中任务选择T01~T11代表的11个分类任务，$f_{features/disk}$代表一张平滑且有盘状结构的频率，$f_{edge\text{-}on,no}$代表一张图像没有侧向边缘的频率，$f_{spiral,yes}$代表一张图像是旋涡星系的频率，以此类推；SDSS DR16主要物理限制中，*g.lnLDeV_g*中*g.*是Galaxy库的一个代称，*lnLDeV_g*代表的是g波段崩解曲线拟合的可能性， *lnLExp_r*代表的是r波段指数拟合的可能性。N_{sample}为样本总数。

2.2 星系图像预处理

卷积神经网络对尺寸小的数据学习能力更强，且训练速度快[17][18]。为了减小图像中存在的不必要相邻信息对实验结果的影响，我们首先对星系数据进行了剪裁处理，之后进行了下采样。以透镜星系为例，如图2所示424 x 424pixel的图像被剪裁成164 x 164pixel，之后将图像下采样到80 x 80pixel大小。

图像在相机捕捉、图像信息传输、数字图像转化过程中等都会存在噪声干扰，噪声的叠加会严重影响图像质量，进而导致图像的本质特征发生改变。对星系形态进行分类时，保存图像中星系的外形轮廓和纹理信息至关重要，所以本文对图像采用边缘导向的非局部均值去噪方法[20]。首先，对图像采用二阶差分Sobel算子抽取边缘；其次，将边缘信息与原有的噪声图像共同构建一个非局部协同滤波框架；最后，将边缘信息参与噪声图像的修复。去噪效果如图3所示，可以发现去噪之后星系周围的噪声点被去掉，且图像有了更多、更明显的边缘纹理信息。

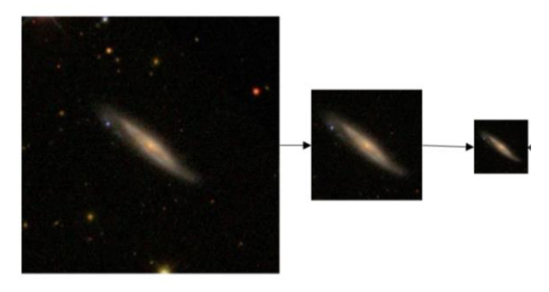


图2 星系剪裁及下采样

Fig.2 Galaxy image clipping and down sampling

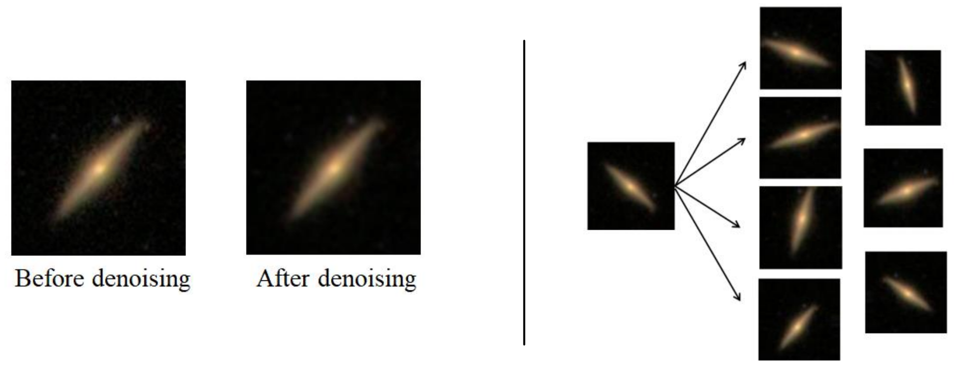


图3　图像去噪效果展示

Fig.3　Image denoising effect display

图4　数据增强效果展示

Fig.4　Data enhancement results display

数据集中不规则星系和透镜星系类型的数量相对较少，数据集的类间比例失衡会影响模型的可靠性。所以本文采用数据增强的方法增加不规则星系和透镜星系的个数。数据增强效果如图3所示，数据增强方式如下[19]：

旋转：星系图像具有旋转不变性，利用图像的这一性质对图像进行随机旋转，旋转范围设置为30°；

缩放：缩放范围为0.7-1.3倍；

翻转：沿着垂直轴和水平轴随机翻转每个图像；

平移：图像中的对象可能不在帧的中心，并且在不同方向上会有偏移。我们对每幅图像进行了水平和垂直随机平移，平移范围为0-10像素。

3 分类网络介绍

3.1 GMC-net网络构架

如图4所示，典型的ConvNet[21]由输入层、卷积层、池化层、全连接层以及最后的输出层构成。输入层主要是把初始化数据做预处理；卷积层主要进行特征提取；池化层主要进行特征压缩，减小过拟合；全连接层主要起到“分类器”的作用。

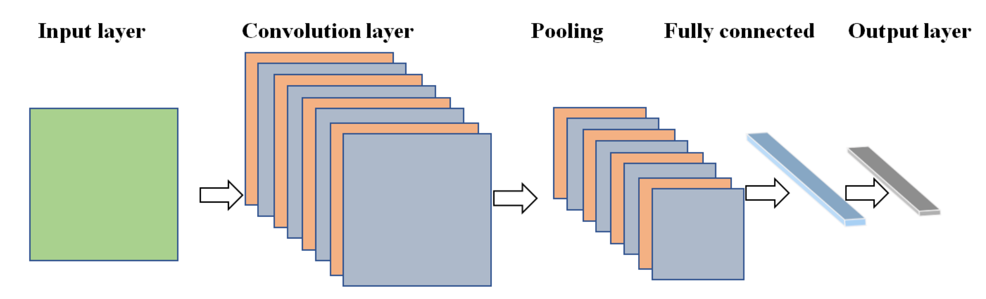


图5 卷积神经网络通用结构

Fig.5 General structure of convolutional neural network

本文受Lenet5网络参数量少易训练优点的启发，结合不同激活函数和BN层的特点，搭建了GMC-net网络。该网络不仅训练的参数量少，还因BN层的加入大大加快了网络的收敛速度，并获得了很好的分类准确率。

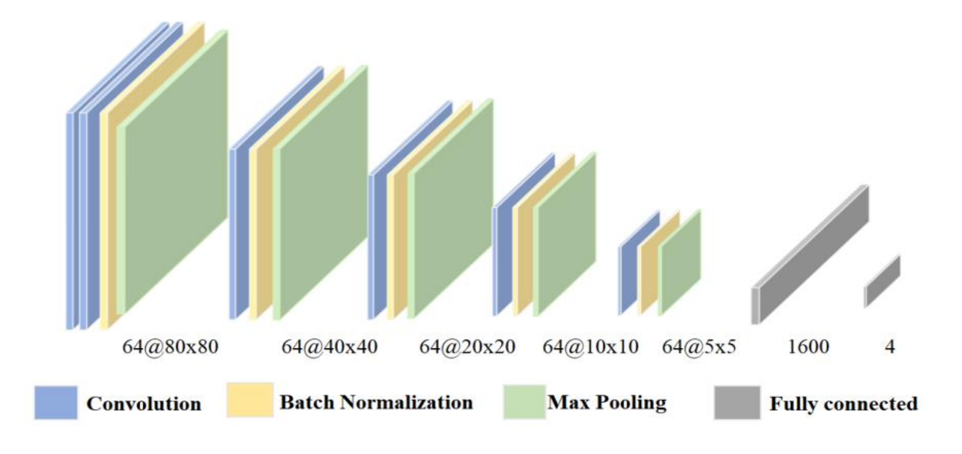


图6　GMC-net整体构架

Fig.6 Overall framework of GMC-net

图6是GMC-net整体构架图，该网络有一个输入层、五个卷积层及一个全连接层和一个输出层，表2是对GMC-net构架中各个层的参数设置总结。GMC-net网络的每一层卷积层后都有一个BN层和最大池化层。BN层可以加快收敛速度和训练速度，池化层对卷积得到的特征进行特征压缩来减小过拟合。此外，GMC-net网络采用不同的激活函数相互协调：为更好输入到下一层前两层使用双曲正切激活函数(Tanh)[22]；为使模型的收敛速度稳定、计算速度更快，中间第三、四卷积层使用修正线性单元ReLU (Rectified linear unit)[23]激活函数；为抑制神经元死亡第五层卷积层采用Leaky ReLU激活函数。经过第五层卷积层之后的特征被Flatten()函数展为一维数组，并输入第一层全连接层，在此所用激活函数为ReLU激活函数，输出为1600。输出层设置为4向分类，所用的激活函数为softmax。

表 2 GMC-net 体系结构概述

Tab.2 Overview of GMC-net architecture

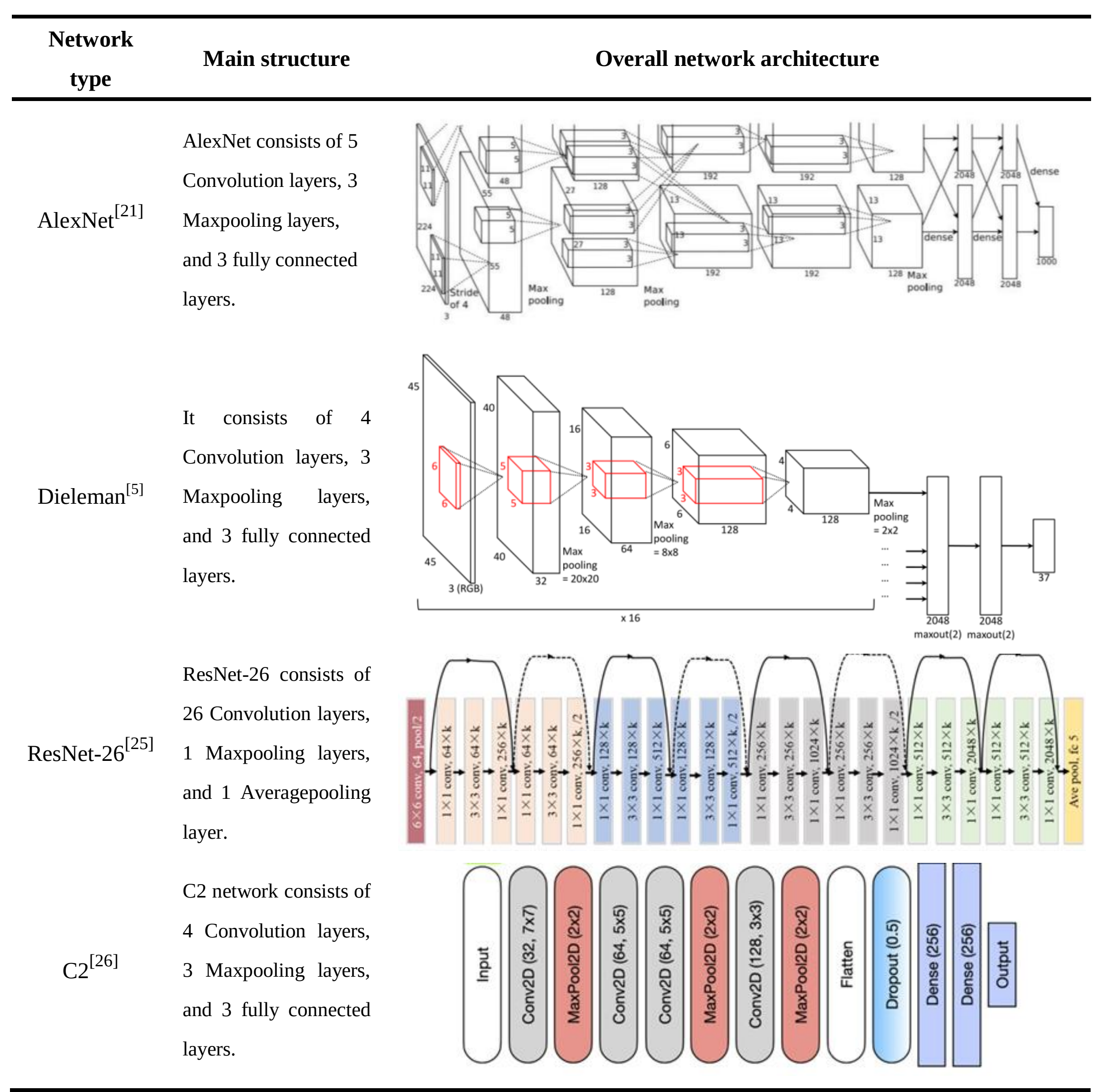
	Filters	Filter size	Padding	Activation function	Type
Conv_1	64	3 x 3	Same	Tanh	2D
BN_1	-	-	-	-	-
Pooling_1	-	-	-	-	Max pooling
Conv_2	64	5 x 5	Same	Tanh	2D
BN_2	-	-	-	-	-
Pooling_2	-	-	-	-	Max pooling
Conv_3	64	5 x 5	Same	ReLU	2D
BN_3	-	-	-	-	-
Pooling_3	-	-	-	-	Max pooling
Conv_4	64	7 x 7	Same	ReLU	2D
BN_4	-	-	-	-	-
Pooling_4	-	-	-	-	Max pooling
Conv_5	64	7 x 7	Same	Leaky ReLU(alpha=0.01)	2D
BN_5	-	-	-	-	-
Pooling_5	-	-	-	-	Max pooling
Fully_1	1600	-	-	ReLU	-
Output	4	-	-	Softmax	-

3.2 其他分类网络介绍

本次研究还用了Krizhevsky等人提出的AlexNet网络[21]、基于Dieleman等[5]提出的卷积神经网络、戴佳明等[25]提出的ResNet-26网络以及Cavanagh等[26]针对星系形态分类提出的C2分类网络。

表 3 其他分类网络简介

Tab.3 Introduction to other classified networks



Network type	Main structure	Overall network architecture
AlexNet[21]	AlexNet consists of 5 Convolution layers, 3 Maxpooling layers, and 3 fully connected layers.	
Dieleman[5]	It consists of 4 Convolution layers, 3 Maxpooling layers, and 3 fully connected layers.	
ResNet-26[25]	ResNet-26 consists of 26 Convolution layers, 1 Maxpooling layers, and 1 Averagepooling layer.	
C2[26]	C2 network consists of 4 Convolution layers, 3 Maxpooling layers, and 3 fully connected layers.	

表3中分别对AlexNet网络、Dieleman网络、ResNet-26网络以及C2网络的整体结构结构进行了简单介绍，其构架图中可以清楚的看到整体网络的层数、每个层所在的位置、每一层滤波器数量及大小的设置参数、所用池化层的池化方式以及Dropout层的丢弃率等。此外四个分类网络所有的卷积层都是采用的ReLU激活函数。

4 实验结果分析及讨论

在本节中，我们首先对评估模型的性能指标进行了介绍，之后用不同网络对星系数据进行分类并与类似的研究进行了对比。

4.1评价指标参数介绍

1、通过混淆矩阵(如表4所示)，可以求得衡量分类模型的性能指标：准确率、精确率、召回率以及F1-score调和值。

表 4　混淆矩阵

Tab.4　Confusion matrix

		Predicted value	
		True	**False**
Actual value	**True**	TP	FN
	False	FP	TN

TP (True Positive)：把正样本成功预测为正；**TN** (True Negative)：把负样本成功预测为负；

FP (False Positive)：把负样本错误地预测为正；**FN** (False Negative)：把正样本错误的预测为负。

准确率(Accuracy)反映的是分类模型所有判断正确的结果占总观测值的比重；精确率(Precision)是在模型预测是 Positive 的所有结果中，模型预测正确的比重；召回率 (Recall)是在真实值是 Positive 的所有结果中，模型预测正确的比重；F1-score 是精确率和召回率的调和平均数。准确率、精确率、召回率及 F1-score 的计算公式分别如公式(1)、(2)、(3)、(4)所示：

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\text{-}score = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

4.2 训练和验证

本文的所有程序都是python程序，运行在2.80 Ghz Intel(R)Core(TM)i9-10900F CPU、16GB内存和64位Windows系统的桌面上，并使用RTX 2070 super GPU加速计算。在模型训练过程中，由于batch size的大小取决于数据集大小以及GPU的处理能力，综合考虑我们将batch size设置为64。

本次研究首先对综合数据集中(Galaxy zoo2 、SDSS DR16、EFIGI目录)的四种不同形态星系进行了分类测试，在模型训练开始前，首先将数据集分为了按照7.5：2.5分为训练集和验证集，并对两者分别进行了数据增强，最终数据集构成如表5所示。

表 5　数据集信息

Tab.5 Dataset information

	Data set 1		**Data set 2**	
	Training set	Test set	Training set	Test set
Spiral	3869	1289	1005	300
Elliptical	4956	1651	1012	300
Lenticular	4067	1355	1017	321
Irregular	3680	1227	1003	305
Total data	16572	5522	4037	1226

表5中的数据集1(Data set 1)是来自SDSS DR16、Galaxy Zoo2和EFIGI目录三方的综合数据集，由于表1中透镜星系和不规则星系数量相较于其他两类较少，为减少类间比例失衡问题对分类模型的影响，在此对透镜星系和不规则星系进行了数据增强。数据集2(Data set 2)是EFIGI目录单独构成的数据集，原始数据为920张旋涡星系、289张椭圆星系、531张透镜星系以及248张不规则星系。为保持各类形态星系类间比例均衡，对每类星系也进行了不同程度的数据增强。最终数据集1中16572张图像作为训练集，5522张图像作为测试集；数据集2中4037张图像作为训练集，1226张图像作为测试集。

在训练及验证过程中，如图7所示我们对GMC_net网络、C2网络、AlexNet网络、Dieleman提出的分类网络以及ResNet-26分类网络的可训练参数量进行了统计。

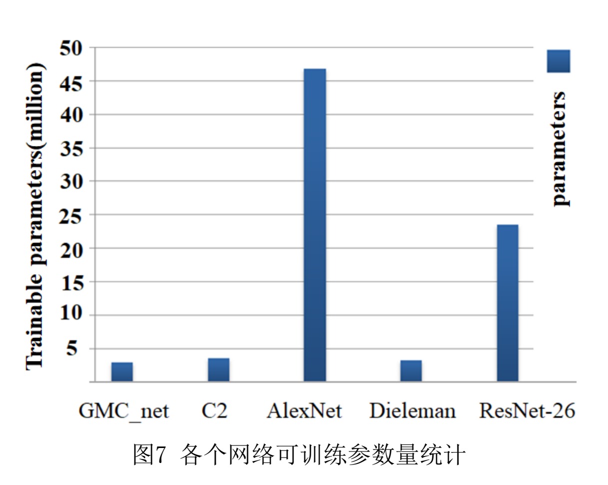


图7 各个网络可训练参数量统计

Fig 7 Statistics of trainable parameters of each network

网络可训练参数量反映了该网络计算过程中的复杂程度，是决定模型的训练速度的重要因素。参数量越大说明网络越复杂，同一设备下训练该网络所消耗的时间越多，且越复杂的网络对计算机计算性能的要求越高。从图中可以看出AlexNet网络和ResNet-26网络的可训练参数远远大于其他三个网络。其中Dieleman的可训练参数约为362万，C2网络的可训练参数约为357万，GMC_net网络的约为293万。从可训练参数量来看GMC_net网络是最少的，其在训练速度上占了很大优势。

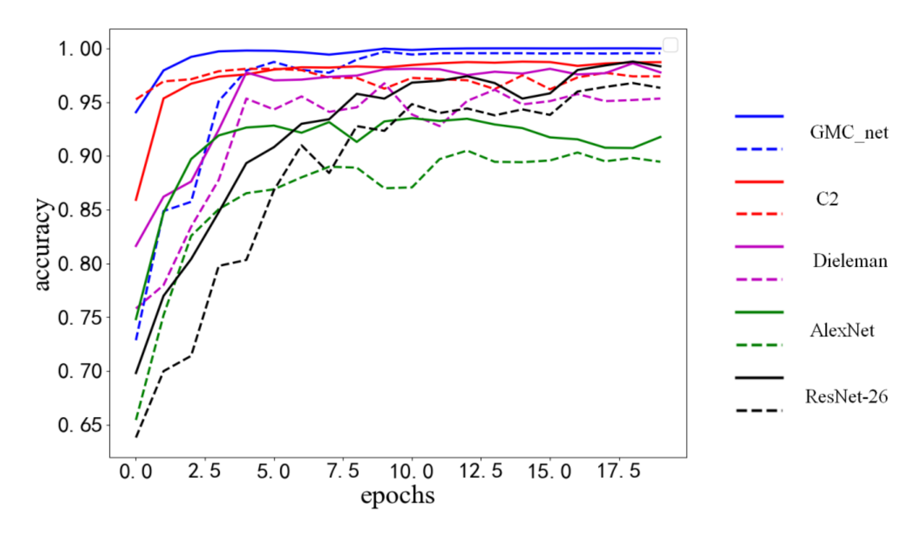


图8 准确率与训练次数关系曲线图

Fig 8 Graph of relationship between accuracy and epochs

注：图中的实线为训练集准确率与训练次数的变化关系曲线，虚线为验证集准确率与训练次数的变化关系曲线

图8显示了五种CNN架构的在训练时训练集准确率、验证集准确率随训练次数的变化趋势(在此所有的权重和偏差在训练开始时都是随机的)，在此我们展示了20个epochs与准确率的关系图。从图8中可以发现，五个分类网络的准确率都呈现出迅速上升之后趋于稳定的趋势。其中AlexNet网络在训练10次左右开始趋于收敛，最终训练集的最高准确率为92.3%，验证集的最高准确率为90.0%；Dieleman网络在训练7次左右开始趋于稳定，训练集准确率最高为96.3%，验证集准确率最高为95.2%；ResNet-26网络在训练16次左右开始趋于稳定，其收

敛速度相对较慢，训练集最高准确率为98.2%，验证集最高准确率为97.8%；C2网络在训练6次左右开始趋于稳定，训练集准确率最高为98.5%，验证集准确率最高为97.9%；GMC_net网络在训练4次左右开始趋于稳定，训练集准确率最佳是为99.53%，验证集准确率最佳为99.18%；从图8可以看出，GMC_net网络在训练过程中准确率是最高的。在各个网络训练最佳情况下，耗时最多的是AlexNet和ResNet-26网络，耗时最少的是GMC_net网络。

综上，GMC_net网络的可训练参数最少，且训练过程中GMC_net其训练集和验证集的准确率均能保持稳定且高于其他四个网络，在收敛速度上超过了其他四个网络，总体来看GMC_net表现最好。

4.3 不同分类方法的分类结果对比

表6是GMC_net对数据集1中验证集测试得到的混淆矩阵，通过混淆矩阵可以计算得到相应的准确率、精确率、召回率以及F1-score。

表 6　数据集 1 验证集分类测试的混淆矩阵

Tab.6 Confusion matrix of verification set classification test in data set 1

		Predicted value				**Recall**
		Spiral	Elliptical	Lenticular	Irregular	
Actual value	Spiral	1269	12	7	1	98.44%
	Elliptical	12	1635	4	0	99.03%
	Lenticular	5	10	1340	0	98.89%
	Irregular	5	3	0	1219	99.34%
Precision		98.29%	98.49%	99.18%	99.91%	
F1-score		98.36%	98.75%	99.03%	99.62%	
Accuracy		98.93%				

由表6可以得出，本次实验最终对旋涡星系的分类精确率为98.29%，其召回率为98.44% ，所得的F1-score值为98.36%；椭圆星系的分类精确率为98.49%，召回率为99.03%，其F1-score值为98.75%；透镜星系的分类精确率为99.18%，召回率为98.89%，其F1-score值为99.03%；不规则星系的分类精确率为99.91%，召回率为99.34%，其F1-score值为98.36%；总体分类准确率为98.93%。

表6展示的是数据集1中5522张验证集在五个分类网络的最终分类结果对比，表中的准确率、精确率以及召回率都是各个网络多次重复验证后取得的最佳结果。

表 7　不同网络验证结果对比

Tab.7 Comparison of verification results of different networks

Network	Accuracy	Precision	Recall	F1-score
AlexNet	91.23%	90.15%	92.34%	91.23%
Dieleman	94.92%	95.32%	93.47%	94.38%
ResNet-26	97.82%	98.36%	97.54%	97.94%
C2	98.04%	98.27%	97.96%	98.11%
GMC_net	98.93%	98.96%	98.90%	98.94%

从表7中可以看到，AlexNet和Dieleman在准确率、精确率及召回率上均小于其他网络，两者的F1-score调和值相比于其他网络也偏低；ResNet-26虽然精确率比C2网络要高，但是在

准确率、召回率及F1-score上略低于C2网络，GMC_net在五个网络中获得了最高的准确率，其精确率、召回率以及F1-score调和值与以上所有网络的相比也都为最高。从最终分类效果来看，GMC_net的分类性能优于其他网络。

为进一步证明我们方法的可行性，我们单独针对表1中来自EFIGI目录的星系重新利用GMC_net进行了单独训练分类并与其他研究方法进行了对比。为保持类间比例均衡，我们将来自EFIGI目录的星系扩展为表5中的数据集2。在此根据[7][10]中的数据描述，我们所选的数据集是包含两者所用的样本（所涉及到的样本类型均选取了其所有子类）。

其中[7]提取了星系图像的颜色特征(first three order moments)、纹理特征(灰度共生矩阵，其中包含熵、对比度、相关性、能量等信息)以及其形状特征(contour moments)三种特征，并用二进制正弦余弦算法选择最相关的特征，之后用KNN进行分类测试；[10]是利用四元数极坐标复指数变换矩(qpet)从星系彩色图像中提取色彩特征并进行特征筛选，最终利用极限学习机(EML)来进行分类。

表 8　与其他研究方法的对比结果

Tab.8　Comparison with other studies

Method	Accuracy	Precision	Recall	F1-score
[7]	91.9 %	92.7%	85%	88.68%
[10]	98.71%	98.72%	98.78%	98.74%
GMC_no	99.04%	98.88%	98.76%	98.81%
GMC	99.34%	99.12%	98.86%	98.98%

注：GMC_no与GMC的区别是：GMC_no没有去噪处理这一过程

从表8中可以看出，在都使用EFIGI目录做数据集的前提下，方法[7]对EFIGI目录中的椭圆星系、旋涡星系、透镜星系以及不规则星系进行分类，效果最好的为分类精确率为92.7%，其F1-score的值为88.68%；方法[10]对四类星系进行分类的最佳结果总体召回率为98.78%，其F1-score的值为98.74%；未进行去噪处理时，GMC_no的召回率低于方案[10]的召回率，去噪之后，GMC对EFIGI目录中椭圆星系、旋涡星系、透镜星系以及不规则星系分类的总体分类准确率、精确率、召回率以及F1-score值均提升，且比方案[7][10]得到准确率、精确率、召回率及调和值都要高。

其次，方案[7][10]以上两种方法一方面在特征选择、分类器选择上有很大困难，且其处理、运算过程比较复杂；另一方面[7][10]存在星系分类样本类间比例严重失衡的问题，这极大可能导致实验结果不能真实反映真实分布，甚至直接估计出现很大误导。而本文所采用的方法在前期对图像进行了预处理，一是用非局部均值去噪减少了噪声对图像的影响，二是我们对不同形态的星系分别进行了数据增强，减小了由于样本量小、样本类间比例分布不均对实验结果产生的影响，最后采用GMC_net分类网络完美避开了图像特征提取、选择及分类器的选择难题，所以综合来看我们的分类方法是非常可行的。

4.4　GMC_net网络卷积特征可视化

本次研究最后利用Grad-CAM[29]技术对GMC_net卷积特征进行了可视化解释，gard-CAM可以将热力图与原图结合的方式将各类形态星系经过卷积之后的特征进行展示，gard-CAM图可以反映卷积神经网络对预测输出的贡献分布，分数越高的地方表示原始图像对应区域对网络的响应越高、贡献越大。

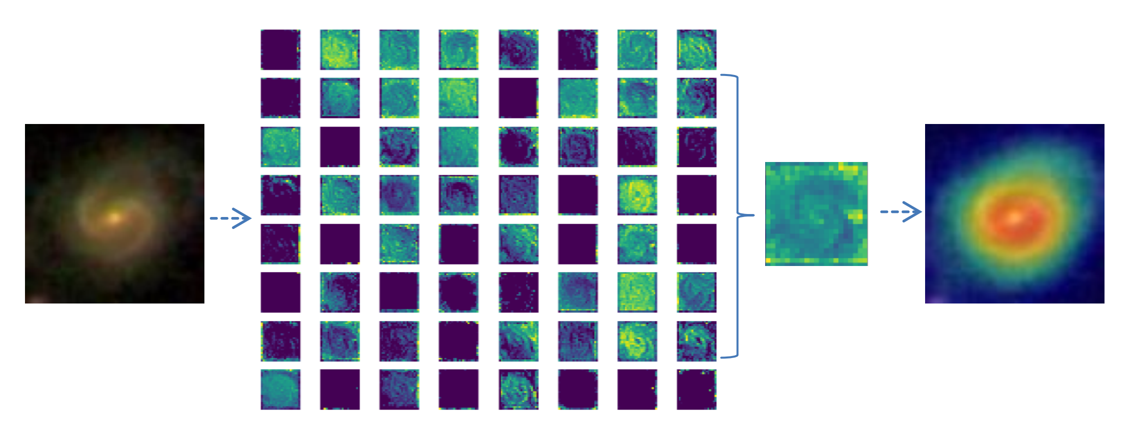


图10 旋涡星系经GMC_net卷积后特征可视化

Fig. 10 Spiral galaxies passing through GMC_ Net convolution feature visualization

注：每幅子图从左往右依次是星系原图、经过GMC_net第四层卷积之后的map图、map合并图、gard-CAM可视化图像

GMC_net不同卷积层所提取特征不同，最开始提取的星系边缘、角落等，之后边缘检测提取简单形状。在高层中，特征图利用高级特征的组合来识别抽象斑点。以旋涡星系为例，例如在第四卷积层中，图10所示map合并图中每个要素图的可区分性更强，这正是分类模型所期望的。利用gard-CAM对经过四层卷积的特征进行了可视化，图中清楚地展现其核心中间的突起及涡旋星系旋的臂状结构，特征贡献度由内向外螺旋递减，进一步清楚地展现了GMC_net在星系形态在星系轮廓特征、纹理特征提取及处理方面的高性能。

5 总结与展望

星系的形态与星系的形成、演化有着密切的联系，是探究星系物理的重要参数。目前对于星系形态分类研究领域依然存在分类类别少、图像特征选择困难、各类形态星系样本分布不均、分类的准确率较低等问题。针对以上问题，本文提出了一种基于卷积神经网络的星系形态分类方法GMC，实现了对旋涡星系、椭圆星系、透镜星系已经不规则星系四种形态的高效分类。本次研究中，我们首先对星系图像进行剪切、下采样、去噪、数据增强一系列处理，保证了个样本的多样性、均衡性，减小了图像噪声和样本类间比例失衡对分类模型的影响；其次，我们构建了一个针对星系形态分类卷积神经网络—GMC-net，此网络可以自动提取星系图像的特征，并根据其形态进行自动分类，避开了特征提取、选择以及分类器选择的难题。利用GMC方法对综合数据集(SDSS DR16、Galaxy Zoo2、EFIGI目录组合)中不同形态的星系进行了分类，从实验分类结果来看，旋涡星系、椭圆星系、透镜星系以及不规则外形星系分类精确率分别为98.29%、98.49%、99.18%、99.91%，召回率分别为98.44%、99.03%、98.89%、99.34%；对来自EFIGI目录中四种形态星系的分类平均分类准确率也达到了99.34%。实验结果表明GMC相较于其他分类方法表现更好，可以更有效地用于星系的形态分类。

本文虽然在一定程度上推动了星系形态分类问题的解决，取得了相应的进展，然而仍然存在一些不足之处有待进一步探索：

首先在数据上为保证所选样本更准确，本文在Galaxy Zoo2中所选择的阈值都是偏大一些的，对该数据集应用的还是不够充分；其次在SDSS DR16中由于对不规则星系的物理参数还未有人统计研究，在此未直接从DR16中得到不规则星系。星系形态分类无疑是需要大量的样本量，获取数据的方式也很多，未来在数据方面可以从数据库利用率以及五波段测光数据应用等方面进行研究。

其次本文所构建的GMC_net网络可自动提取星系形态特征，并自动对星系形态分类。从分类结果来看分类准确率很好，但其中透镜星系、椭圆星系及涡旋星系错分的图像相对多一

点，且对错分的样本难以区分。所以在未来分类系统研究中可以尝试构建专家系统与神经网络相结合的混合模型，即神经网络专家系统，以提升模型的分类性能。

致谢：感谢国家自然科学基金委员会-中国科学院天文联合基金（U1931134）、河北省自然科学基金（A2020202001）以及中国科学院天文大科学研究中心 LAMOST 重大成果培育项目对本文工作的大力支持。

参 考 文 献

[1] Lupton R H , Yasuda N . SDSS Imaging Pipelines[J]. 2002.

[2] Scoville N, Aussel H , Brusa M , et al. The Cosmic Evolution Survey (COSMOS) -- Overview[J]. The Astrophysical Journal Supplement Series, 2008, 172(1):1.

[3] Ivezic Z , Tyson J A , Acosta E , et al. LSST: from Science Drivers to Reference Design and Anticipated Data Products[J]. American Astronomical Society, 2008.

[4] Freed M , Lee J . Application of Support Vector Machines to the Classification of Galaxy Morphologies[C]// Fifth International Conference on Computational & Information Sciences. IEEE, 2013.

[5] Sander D, Willett K W, Joni D . Rotation-invariant convolutional neural networks for galaxy morphology prediction[J]. Monthly Notices of the Royal Astronomical Society, 2015(2):2.

[6] Kim E J , Brunner R J . Star-galaxy Classification Using Deep Convolutional Neural Networks[J]. Monthly Notices of the Royal Astronomical Society, 2016.

[7] Selim I M , Mohamed A . Automated morphological classification of galaxies based on projection gradient nonnegative matrix factorization algorithm[J]. Experimental Astronomy, 2017, 43(2):131-144.

[8] Mittal A , Soorya A , Nagrath P , et al. Data augmentation based morphological classification of galaxies using deep convolutional neural network[J]. Earth Science Informatics, 2019, 13(1).

[9] Mittal M. Morphological classification of galaxies using Conv-nets[J]. Earth Science Informatics, 2020(1):1-10.

[10] Hosny K M, Aziz M, Selim I M, et al. Classification of galaxy color images using quaternion polar complex exponential transform and binary Stochastic Fractal Search[J]. Astronomy and Computing, 2020, 31.

[11] The Sloan Digital Sky Survey: Technical Summary[J]. Astronomical Journal, 2000, 120(3):1579.

[12] Anderson S F, Arns J A, Aubourg E , et al. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems[J]. Astronomical Journal, 2011, 142(3):725-735.

[13] Baillard A, Bertin E, Lapparent V D, et al. The EFIGI catalogue of 4458 nearby galaxies with detailed morphology[J]. Astronomy and Astrophysics, 2011, 532.

[14] Bamford, S. P , Masters, et al. Galaxy zoo 2: Detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey.

[15] Li N, Szalay A. CASJobs: A workflow environment designed for large scientific catalogs. IEEE, 2008.

[16] Flesch E W . Identification confusion and blending concealment in the SDSS-DR16 Quasar catalogues -- 40 new quasars and 82 false quasars identified[J]. 2020.

[17] Yong X, Zhong J . Down-Sampling Face Images and Low-Resolution Face Recognition[C]// International Conference on Innovative Computing Information & Control. IEEE, 2008.

[18] Zhang Y , Zhao D , Zhang J , et al. Interpolation-dependent image downsampling.[J]. IEEE Transactions on Image Processing, 2011, 20(11):3291-3296.

[19] Sander D , Willett K W , Joni D . Rotation-invariant convolutional neural networks for galaxy morphology prediction[J]. Monthly Notices of the Royal Astronomical Society, 2015(2):2.

[20] 傅博,吴越楚,王丽妍,王瑞子.边缘导向的非局部均值图像滤波[J].吉林大学学报(信息科学版),2020,38(06):687-693.

FU Bo，WU Yuechu，WANG Liyan，WANG Ruizi. Edge Map Oriented Non-Local Means Filtering Algorithm[J]. Journal of Jilin University (Information Science Edition) , 2020,38(06):687-693

[21] Fukushima K , Miyake S , Ito T . Neocognitron: A neural network model for a mechanism of visual pattern recognition[J]. Systems Man & Cybernetics IEEE Transactions on, 1982, SMC-13(5):826-834.

[22] Mathias A C, Rech P C. Hopfield neural network: the hyperbolic tangent and the piecewise-linear activation functions.[J]. Neural Networks, 2012, 34(10):42-45.

[23] Pedamonti D. Comparison of non-linear activation functions for deep neural networks on MNIST classification task[J]. 2018.

[24] Technicolor T , Related S , Technicolor T , et al. ImageNet Classification with Deep Convolutional Neural Networks [50].

[25] Dai J M , Tong J . Galaxy Morphology Classification with Deep Convolutional Neural Networks[J]. 2018.

[26] Cavanagh M K , Bekki K , Groves B A . Morphological classification of galaxies with deep learning: comparing 3-way and 4-way CNNs [J]. Monthly Notices of the Royal Astronomical Society, 2021(1):1.

[27] Clevert, Djork-Arné, Unterthiner T, Hochreiter S . Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)[J]. Computer Science, 2015.

[28] Simon M , Rodner E , Denzler J . ImageNet pre-trained models with batch normalization[J]. 2016.

[29] Selvaraju R R, Cogswell M , Das A , et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization[J]. International Journal of Computer Vision, 2020, 128(2):336-359.

A new method for Galaxy morphology classification-GMC

Wang Linqian[1], Qiubo[1], Luo Ali[2], Kong Xiao[2] ,Lu Yakun[1], Guo Xiaoyu[1]

1. Hebei University of Technology, Tianjin, 300400

2. National Astronomical Observatory Chinese Academy of Sciences, Beijing, 100012

Abstract: In the field of astronomy, the classification of galaxies has always been a hot and difficult problem. In recent years, some scholars have applied machine learning to the simple classification task of galaxy morphology, but in the process of classification, there are a series of problems, such as feature selection difficulty, feature omission, classifier selection difficulty and so on. Galaxies can be roughly divided into elliptical galaxies, spiral galaxies, lenticular galaxies and irregular galaxies in visual morphology. In this paper, GMC (Galaxy morphological classification) which is a more accurate classification method is proposed for the photometric images of galaxies in SDSS DR16, Galaxy Zoo2 and EFIGI catalog. Firstly, we cut and denoise the images, and use rotation, translation, scaling and other methods to enhance the data. Finally, we build the GMC-net to classify photometric images. According to the classification results, the classification accuracy of spiral galaxies, elliptical galaxies, lenticular galaxies and irregular galaxies in different databases are 98.29%,98.49%,99.18% and 99.91%, respectively; The average classification accuracy of four different galaxies from the same database EFIGI catalog is 99.34%. The experimental results show that GMC performs better than other classification methods, and can be used to classify galaxies more effectively.

Key words: Galaxy morphology classification; Data enhancement; Convolution neural network;